CLASSIFICATION OF MISSING VALUES HANDLING METHOD DURING DATA MINING: REVIEW

Entin Hartini

Pusat Teknologi dan Keselamatan Reaktor Nuklir - BATAN

ABSTRACT

CLASSIFICATION OF MISSING VALUES HANDLING METHOD DURING DATA MINING: REVIEW. Missing data often occurs in researchs or surveys. Many real datasets or data mining have missing data, thus affecting the quality of the data. There are various causes resulting in incomplete data, such as: manual data entry procedure, incorrect measurement, equipment error, and many others. Any errors causing data missing make it difficult in a data analysis. This is due to the algorithms of data analysis that only work if the data is complete. Missing data analysis may help resolving missing data. Missing data can be replaced with a value based on the possibility of other information available, so that the data set can be analyzed. Many specialists have been working on this issue to present more modern techniques. Many strategies are available for handling the missing data, however investigator has difficulty in finding the right technique in the absence of information about strategy and implementation. The purpose of this research paper is to classify methods of missing data handling based on statistical method and machine learning. Results from this study are classification methods of missing data handling by ignoring technique, model base technique and imputation technique , which are complemented with the advantages and disadvantages of each method.

***Keywords**:* missing value, statistic, machine learning, classification, method

ABSTRAK

***KAJIAN MENGENAI KLASIFIKASI METODE PENANGANAN DATA HILANG SAAT PENGAMBILAN DATA**. Data hilang sering terjadi dalam penelitian atau survei. Banyak kelompok data riil saat pengambilan data yang memiliki data yang hilang, sehingga mempengaruhi kualitas data. Berbagai penyebab yang mengakibatkan ketidaklengkapan data, seperti: prosedur entri data manual, pengukuran yang tidak benar, kesalahan peralatan, dan banyak lainnya. Adanya kesalahan yang mengakibatkan data hilang membuat sulit dalam melakukan analis data. Hal ini disebabkan karena algoritma dari analis data hanya bekerja jika data tersedia lengkap. Analisis data yang hilang dapat membantu mengatasi data yang hilang. Data yang hilang bisa diganti dengan nilai berdasarkan kemungkinan informasi lain yang tersedia, sehingga data set dapat dianalisis. Banyak spesialis yang bekerja pada masalah ini untuk menyajikan teknik yang lebih modern. Strategi yang tersedia untuk menangani data yang hilang cukup banyak, namun demikian kesulitan peneliti adalah dalam menemukan teknik yang tepat dikarenakan tidak adanya informasi tentang strategi dan implementasi. Penelitian ini adalah untuk mengklasifikasikan metode penanganan data yang hilang berdasarkan metode statistik dan machine learning. Hasil kajian ini adalah berupa klasifikasi metode penanganan data yang hilang dengan teknik: ignoring technique, model base technique dan imputation technique, serta keuntungan dan kerugian dari masing-masing metode.*

***Kata kunci:** data hilang, statistika, machine learning, klasifikasi, metode*

INTRODUCTION

Most observation on data set is currently experiencing a problem of missing data. This will lead to an investigation about the mining information, which obtains final conclusions that might be wrong related to the data being studied. Data mining is a process that requires a high availability of large amounts of data, which are needed to be converted into useful information. Data preparation is the main phase in the investigation of the data [1].

The data set contains lost values due to various reasons, such as manual data entry procedures, errors on equipment and during measurement. Three problems associated with missing values are: loss of efficiency, complexity in handling and analyzing the data and bias arising from the difference between missing and incomplete data. The missing data will reduce the precision of calculation because the amount of information is reduced. Therefore a method of handling missing data is required [2].

Previous methods used in dealing with missing data (such as: deleting data that contains incomplete information, or replacing missing values with the approximation of average values) looks very easy to do, but it becomes a problem because these methods will produce biased data model [3].

Many researches have been done by developing inference of missing data [3] and deleting data that contains incomplete information [4]. Research on missing data is performed using hot deck [3-5], imputation regression [3, 5-6], and mean substitution [7-9]. Main approaches for missing data should have good statistical properties as showed by Maximum Likelihood (ML) method having the Expectation Maximum and Multiple imputation (MI). During assessment on Expectation Maximum [10-13], the completion of ML requires algorithms to calculate and maximize the conditional expectation of the log-likelihood function to obtain convergent values. While the completion of the MI method requires prediction model (explicitly) by minimizing and predicting the missing values [13-16].

Data mining algorithms will handle missing data in a very simple way covering techniques of imputation of missing values performed traditionally, such as deleting the data, the mean value imputation, maximum likelihood and other statistical methods. Current research has started investigating the use of machine learning technique as a method of imputation of missing data [1-2, 17]. Missing data handling using machine learning technique has been widely applied. Study on classification of efficient imputation method for analyzing missing values has also been conducted [1-2, 17-19]. K-Nearest Neighbor is normally used in missing data imputation [20-23], while predicting missing attribute values is performed using K- Means Clustering [24-25]. K -NN classifier performs better than K-Means clustering in missing value imputation [26]. Therefore a study of analysis on K-Means algorithm as an imputation method to deal with missing values has been performed [27] followed by survey on the effect of different K -Means Clustering algorithms [28]. On the other

side, algorithm imputation with Fuzzy K-means (FKMI) is accomplished using the euclidean distance function [29,30].

In this study, a grouping method to handle missing data is performed using statistical methods and machine learning technique based on missing data, which are ignoring technique, model base technique and imputation technique. Advantages and disadvantages of these methods are also discussed.

TECHNIQUES FOR MISSING DATA

Missing data is a data having incomplete or missing values. The lost values are caused by various reasons, such as manual data entry procedures, equipment and measurement errors. Three problems associated with missing values are: loss of efficiency, complexity in handling and analyzing the data and bias arised from the difference between missing data and incomplete data. The missing data will reduce accuracy of calculation because the number of information is reduced. Missing value analysis will help resolving problems caused by the absence of data.

There are three mechanisms of missing data [3], which are:

1. Missing completely at Random (MCAR)
   The level of randomness is high in MCAR. If variable A is missing, the data is not dependent on other variable B so that it can not predict the missing variable A from any other variable in data set. So the probability of the missing variable is same for all the missing variables. The advantage of this method is that it is easier for the researchers to estimate and compute the proposed model.
2. Missing at Random (MAR)
   Prediction of the value on missing variable A is dependent on the other variable B in given dataset but not the value of missing data itself. Missing values are dependent on the value of observed information or values in the dataset.
3. Not Missing at Random (NMAR): The missing variables are not random and also can not predicted from other variables in the data set.

Some methods of handling missing value using statistical methods or machine learning are described here. Three approaches to the problem of missing data are:

1. Eliminate all the patterns of the data set containing the lost data. It is very relevant if the data set is small.
2. Replace the missing data (imputation), for example, the average value of the historical data available can use statistical methods or machine learning.
3. Look for a model based on the data to estimate missing values.

Because there are too many methods of handling missing data in the time series data, it is important to study and undertand advantages and disadvantages of each method as also their purposes. Based on study performed, methods of handling missing data can be grouped into two approach, namely using statistics and machine learning. The method of

handling missing values using statistics methods is applied using ignoring technique and model base imputation technique. While the machine learning is imputation technique using K-nearest Neighbours, K-Mean Clustering and Fuzzy C-Means.

MISSING DATA HANDLING METHOD

There are several strategies for missing data handling technique, among them are presented in Figure 1. As previously mentioned, there are two groups of method for handling missing data, which are statistical method

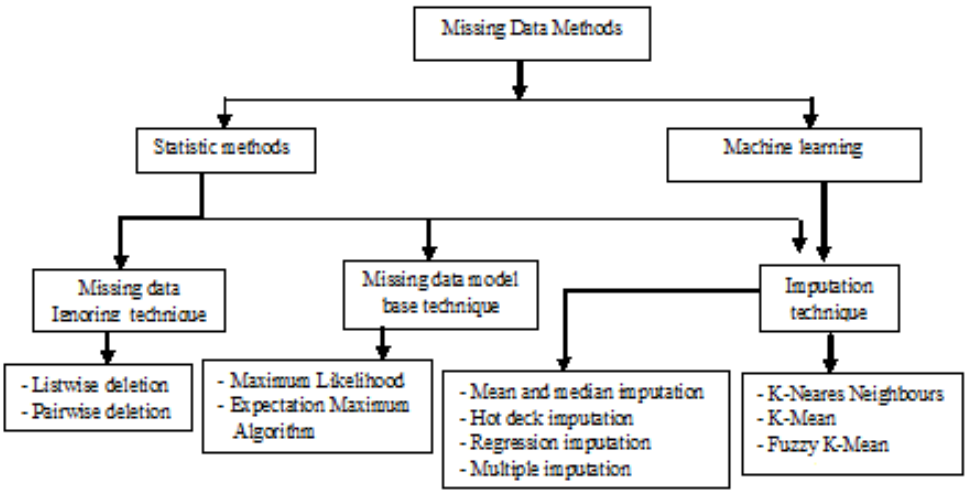

Figure 1. Strategies for missing data handling technique

and machine learning. While missing data techniques can be grouped into three classes, as proposed in reference [17] and described below:

1. Missing Data Ignoring Technique
   Ignoring technique uses listwise and pairwise deletion. Listwise deletion is used if a case has missing data for any of the variables, then essentially that case should be avoided from the analysis. It is typically a default in the statistical package. While pairwise deletion is referred to as the available case method. This technique considers each feature independently. For each feature, all recorded values in each observation are considered and missing data are overlooked [3-4].

2. Missing Data Imputation Methods
   Imputation method involves replacing missing values with estimated values based on some information available in the data set. There are many variation options from that method like mean imputation to some more robust methods based on relationships among attributes as described below:

   - Mean and Median Substitution
     This method is typically used in sample surveys. One instance with missing data (for example, a person that can not be contacted) is replaced by another non sampled instance. The most commonly practiced approach is single impu-

imputation technique. Mean substitution replaces missing values on a variable with the mean value of the observed values. The imputed missing values are dependent on one and only one variable among subjects mean for that variable based on the available data. Mean substitution preserves the mean of a variable distribution, however mean substitution typically distorts other characteristics of a variable distribution. Mean or median substitution of covariates and outcome variables is still frequently used. This method is slightly improved by first stratifying the data into subgroups and using the subgroup average. Median imputation results in the median of the entire data set is same with the case of deletion, but the variability between individual responses is decreased and bias on variances and covariances approach zero [5,7].

- Hot deck

  In the hot deck method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. Hot deck is typically implemented into two stages. In the first stage, the data are partitioned into clusters, followed by the second stage, in which each instance with missing data is associated with one cluster. The complete cases in a cluster are used to fill the missing values. This can be done by calculating the mean or mode of the attribute within a cluster [3].

- Regression Imputation

  Regression imputation is a predictive model in dealing with the imputation of missing data using regression method, in which the values of the features observed and predicted values are then used to fill the missing values [3, 5-6].

- Multiple Imputations

  Multiple imputation methods can generate some complete set data with all the missing imputted values filled by some models such as linear regression model. Variables used to predict missing values must include all variables to be used in parameter estimation based on the analysis models. The overall parameter estimation is the average of all the individual estimates, but the variance of these estimates will reflect the variance in the set data and variance between set data, so there is a calculation of the uncertainty caused by the process of imputation. Thus, multiple imputations (MI) will overcome the limitation of single imputation by pre-

senting a form of additional error based on parameter variations to estimate all imputation errors [13-16].

- K-Nearst Neighbour Imputation (KNNI)

  KNNI method selects the K nearest observation from a series of observations with values known in the attributes to do imputation that will minimize the size of the distance. When the value of K nearest neighbor is found, estimated value is replaced because the value of missing data has to be estimated. Value replacement is calculated depending on the type of data. This method can be used for the data qualitative and quantitative attributes [20-26].

- K-Means

  K-Means is a method to classify or categorize objects based on attributes / features to a number of k groups. K is a positive integer. The grouping is done by minimizing the sum of the squares of the distance between the data and the cluster centroid [26-28]. This technique is a quick and precise way to estimate missing values.

- Fuzzy K-Mean Clustering Imputation (FKMI)

  In FKMI, membership function plays an important role. Fuzzy clustering can provide a better description when each cluster is not well divided. That is the case when an object does not only belong to one particular cluster but also to other clusters. Any object having missing values can cause this object to be inserted into severel clusters. This method will describe the degree of membership of each object on a cluster [29-30].

3. Missing Data Model Base Technique

Model base technique is used to estimate model parameters for all data sets. The procedure is by using a variant to estimate the missing data set parameters.

- Maximum Likelihood

  The maximum likelihood technique is used to estimate model parameters for all data sets. Distribution of the data set is assumed to maximize likelihood (MLE). Estimation of variables can be obtained as the solution of the equation likelihood of observed data. The roots of this equation will globally maximize the likelihood of observed data so that it becomes consistent. This technique is used to seek an estimation of the covariance matrix for the variables in the model of large samples. It is possible to calculate the iterative MLE maximization of variables using the Newton-

Raphson procedure, Fisher score or Quasi-Newton method [10-11].

- Expectation Maximization (EM) Algorithm

  EM algorithm consists of two phases, namely a step-E and step-M. Step-E requires an algorithm to calculate a conditional expectation of log-likelihood function and procedure to maximize a conditional expectation. The missing value of observed data variable is substituted by mead and constional covariance. In the step-M, ML estimation of the average matrix vector and covariance is obtained as if there is no missing value. The results of covariance matrix and the regression coefficients from step-M is used to obtain new estimates of the missing values. Iteration is repeated until the missing values are obtained. This method requires a large sample size and the mechanism of data missing at random (MAR) [12-13].

An overview of the overall method of handling missing values is presented in Table 1. The advantages and disadvantages of these methods of handling missing values are presented in Table 2.

Table 1: Overview of the overall method of handling missing values

Method	Explanation
Listwise deletion	Deletion of all cases that contain missing values. Loss of information is quite high.
Pairwise deletion.	Deletion of records only from columns that contain missing values. Less missing of information by keeping all the available values
Mean and Median Substitution	Replacing missing values with the mean of the data. Imputation will generate mean and standard deviation higher than the original data.
Hot deck	The missing value will be replaced with the observed response of the unit is "identic".
Regression Imputation	Replacing missing values with the values estimated from observed values. The regression equation is: $Y = a + b X$
Multiple Imputations	Completing the limitations of single imputation and replacing any missing items with two or more acceptable values.
K-Nearst Neighbor Imputation (KNNI)	This method uses K-NN algorithm for estimating and replacing the lost data and can estimate both quantitative and qualitative attributes of that attribute.
K-Means Imputation	Use of algorithm called nearest neighbor to replace missing values in the same way as KNNI
Fuzzy K-Mean Imputation (FKMI)	For each data attribute that has not been replaced by FKMI, it is resolved on the basis of degree of membership and values of cluster centroid.
Maximum Likelihood (ML)	This is a parameter estimation method of observation given statistical model. Parameter values are obtained by maximizing the likelihood of a parameter of observation
Expectation Maximization (EM)	Iterative methods using ML consisting of two steps: Expectations (E step) and Maximization (M-step) iteration until the algorithm converges.

Table 2. Advantages and disadvantages of method of handling missing values

Method	Technique	Advantages	Disadvantages
Statistical Methods	1. Ignoring technique - Listwise deletion - Pairwise deletion	The easiest way to do, throw or not to include missing data in the calculation	This method will have a standard deviation that is quite large when there are a lot of missing data, resulting in reduced accuracy of the estimate.
	2. Model base technique - Maximum Likelihood	No need to do the assessment of the missing variables and it is more rapidly to be convergent.	Necessary to find derivatives to two of the distribution function. Before using this method, it needs to test homogenity variants and necessary limit to the number of missing data is allowed.
	- Expectation Maximum (EM)	EM easy to make the program and requires little storage memory. Besides the EM algorithm does not need to calculate the second derivative matrix	This algorithm optimization alternately performs against some required variables in the model. Parameter estimation accuracy depends on the assumed distribution
	3. Imputation technique - Mean Substitusion	Not necessary to build a model of the data. Filling in missing data value with the expected value so as to have a high degree of stability.	Variance obtained by this method does not correspond to the actual data, which will cause estimation error which is always lower than the actual.
	- Hot deck	Completion of mean substitution method, is more stable than the mean substitution.	If there are a lot of missing data, it will be resulting in charging its worth over and over, so that the results of the estimate will be biased
	- Regresion	Produce small standard deviation.	Prediction missing data is done through the regression model
	- Multiple Imputation	Estimated overall parameter is the average of all estimates of the individual, so that generate a small standard deviation.	The variance of these estimates will reflect the variance in the data set and variant data between specified sets, so there is a calculation of the uncertainty caused by the process of imputation.
Machine Learning	Imputation technique - K-Nearest Neighbours	It can predict both quantitative and qualitative data. Easily handle multiple missing values.	It estimates the most similar values. Time consuming process because it searches all instances of similar data set.
	- K-Mean	Easily classify or group the data. Fast and accurate	Difficult to predict K value. It didn't work well with the cluster of global data, different size and density.
	- Fuzzy K-Mean	Estimating missing values. Can be used for quantitative and qualitative data	The more variations of values that attribute, then standard deviation obtained will be even greater.

Technique base model using maximum expectation algorithm is used the best for data having a distribution function in the form of a model equation, such as multivariate normal distribution, mixture Gaus and other types of distribution. For data sets that do not require the model, then the imputation technique is recommended.

Handling of missing data using statistical methods can be performed for data sets with small sample quantities. While for the data set with a very large number of samples, it is more advisable to use imputation technichque with machine learning.

Imputation technique of machine learning is easier to use for real-time dataset. To see which method is more efficient and profitable, a standard error criterion may be used, such as Root Mean Square Error (RMSE). If the value of the RMSE technique is small, then the handling of missing data become more efficient. The author has performed a comparison of some techniques with application to the evaluation of maintenance history data of the primary coolant system with multiple components. Handling techniques of missing values, which were analyzed, are: listwise deletion, substitution mean imputation, and maximum expectation. The results obtained are that the maximum expectation technique possessed the smallest error standard. As for the comparison, listwise deletion, mean substitution and machine learning technique to K-nearest Neighbours imputation (KNNI) produce a small RMSE on KNNI technique. Further research can be done to compare the methods of handling missing data by using statistical methods and machine learning in overall on real-time data or other implementations.

CONCLUSION

This research mainly focuses on the study of methods of handling missing data in data mining. In this study, the overall view of the method of handling missing data with statistical methods and machine learning is discussed. Imputation technique is widely used to fill missing values of various types of data sets. In this way, various proposed strategies can be presented for handling missing values in the data set. The use of imputation technique is more practical, because it does not need the model establishment such as technique model example of Expectation Maximum algorithm. The precision of Expectation Maximum algorithm is better than the imputation technique using statistical methods. Multiple imputation is very suitable to predict, but in some cases the algorithm becomes longer in the calculation process when a prediction is to be calculated in real time. While the precision using the machine learning imputation method is better than the imputation using statistical methods. Further research might be proposed to perform a comparison technique imputation method and the base model of the dataset using the software.

ACKNOWLEDGEMENT

The author would like to thank Dr Syaiful Bakhri and Dr Hendry Julwan Purba, who had given advices and guidances during

this research project. This research has been funded by BATAN financial budget of the year 2016.

REFERENCES

1. BHAVISHA SUTHAR, HEMANT PATEL, ANKUR GOSWAMI, "A Survey: Classification of imputation methods in data mining", IJETAE, 2 (1), 309-312, (2012).
2. UMATHE, VAISHALI H G.C., "A Review on Incomplate Data And Clustering", Int. J. Compt.Sci.Inf. Technol, 6 (2);1225-7, (2015).
3. MARINA SOLEY-BORI , "Dealing with missing data: Key assumptions and methods for applied analysis", Technical Report , 4, (2013).
4. AMANDA N. BARALDI, CRAIG K. ENDERS, "An introduction to Modern Missing Data Analyses", Journal of Scool Psycology, Elsevier, 48, 5 – 37, (2010).
5. ARUMUGA NAINAR S A, "Comparative Study of Missing Value Imputation Methods on Time Series Data", Int.J.Technol. Innov.Res, 14, 1-8, (2015).
6. NOOKHONG J., KAEWRATTANAPAT N., "Efficiency Comparison of Data Mining Techniques for Missing Value Imputation", J. Ind. Intellegent Inf, 3(4), 305-309, (2015).
7. SOMASUNDARAM R.S., NEUDUNCHEZHIAN R, "Missing Value Imputation Using Refined Mean Substitution", Int. J. Comput. Sci, 9(4), 306 -313, (2012).
8. HARTINI E, "Efficiency Comparison of Method of Handling Missing Value In Data Evaluation System or Component", SENTEN , (2016).
9. HARTINI E, "Implementation of Missing Value Handling Method for Maintenance History Data Evaluation of System/Component", Journal Tri Dasa Mega, (2017).
10. PAUL D. ALLISON, "Handling Missing Data by Maximum Likelihood", SAS Global Forum, Statistics and Data Analysis,, Paper 312-2012, Statistical Horizons, Haverford, PA, USA, (2012).
11. CRAIG K. ENDERS, "Applied Missing Data Analysis", Series Editor's Note by Todd D. Little The Guilford Press, Inc. 72 Spring Street, New York London, (2010).
12. N. BALAKRISHNAN, DEBASIS KUNDU, "Hybrid Censoring: "Models, Inferential Results and Applications", Journal Computational Statistics and Data Analysis, 57, 166-209, (2013).
13. ANDERS HANSSON, RAGNAR WALLIN, "Maximum Likelihood Estimation of Gaussian Models with Missing Data Eight Equivalent Formulations", Journal Automatica, Elsevier, 48, 1955 -1962, (2012).
14. DONG Y. PENG C.J, "Principled missing Data Method for Researchers", Spinger Plus, 2004: 1-17, (2013).

15. KE-HAI YUAN, VICTORIA SAVALEI, "Consistency, bias and efficiency of the normal Distribution-Based MLE: The Role of Auxiliary Variables" , Journal of Multivariate Analysis 124, 353 – 370, (2014).
16. GRAHAM, J. W., "Missing data analysis: Making it work in the real world. Annual review of psychology", 60:549-576, (2009).
17. S.KANCHANA, ANTONY SELVADOSS THANAMANI, "Classification of Efficient Imputation Method for Analyzing Missing Values", International Journal of Computer Trends and Technology (IJCTT), 2, 193-195, (2014).
18. N. A. ZAINURI, A. A. JEMAIN,N MUDA, "A Comparison of Various Imputation Methods for Missing Values in Air Quality Data", Sains Malaysiana, 44 (3), 449–456 , (2015).
19. MINAKSHI, RAJAN VOHRA G., "Missing Value Imputation in Multi Attribut Data Set", Int.J. Comput. Sci.Inf. Technol, 5(4), 5315-5321, (2014).
20. ARSLAN I.B.A., "Novel Hybrid Approach to Estimating Missing Value in Databases Using K-Nearest Neighbors and Neural Networks", Int. J. Innov. Comput. Inf. Control, 8(7), 4705-4717, (2012).
21. MALARVIZHI M.R., THANAMANI A.S., "K-Nearest Neighbor in Missing Data Imputation", Int. J. Eng. Res. Dev, 5(1), 5-7, (2012).
22. Wael M.Khedr A.M.E., "Pattern Classification for Incomplate Data Using PPCA and KNN", J.Eng, Trends Comput. Inf. Sci, 4(8), 628-632, (2013).
23. SATISH GAJAWADA, DURGA TOSHNIWAL, "Missing Value Imputation Method Based on Clustering and Nearest Neighbours", International Journal of Future Computer and Communication, 1(2), 206-208, (2012).
24. SUGUNA N., THANUSHKODI K.G., "Predicting Missing Attribute Values Using K- Means Clustering", J. Comput. Sci, 7(2), 216-224, (2011).
25. SWEETY BAIWAL, ABHISHEK RAGHUVANSHI, "Imputation of Missing Values using Association Rule Mining & K-Mean Clustering", International Journal of Scientific Development and Research (IJSDR), 1(8), 340-344 , (2016).
26. MALARVIZHI M.R., THANAMANI A.S., "K-NN Clasifier Perform Better Than K-Means Clustering in Missing Value Imputation", J. Comput. Eng, 6 (5), 12-5, (2012).
27. B. MEHALA, K. VIVEKANANDAN, AND P. R. J. THANGAIAH, "An Analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values", Asian Journal of Information Technology, 9, 434-441, (2008).
28. FATEMEH AHMADI BAKHSH, KEIVAN MAGHOOLI, "Missing Data Analysis: A Survey on the Effect of Different K-Means Clustering Algo-

rithms", American Journal of Signal Processing, 4(3): 65-70, (2014).

29. C. CHANG, J. LAI et al., "A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement", Journal of Information Science and Engineering, 27, 995-1009, (2011).
30. SCHMITT P., MANDEL J., GUEDJ M., "A Comparison of Six Methods for Missing Data Imputation", J Biomet Biostat, 6(1), 1-6, (2015).